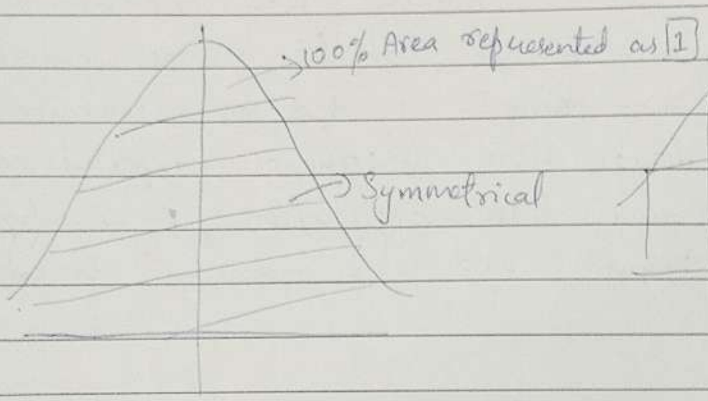
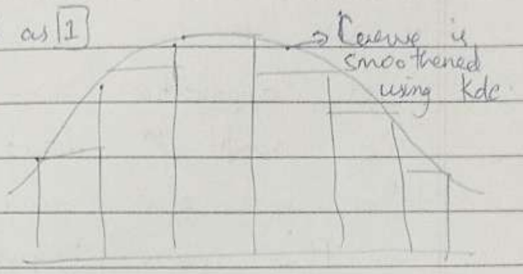
RANKA
DATE / /
PAGE

STATS CLASS 3

1. Normal distribution
2. Standard Normal distribution
3. Z-score.

* Gaussian / Normal distribution



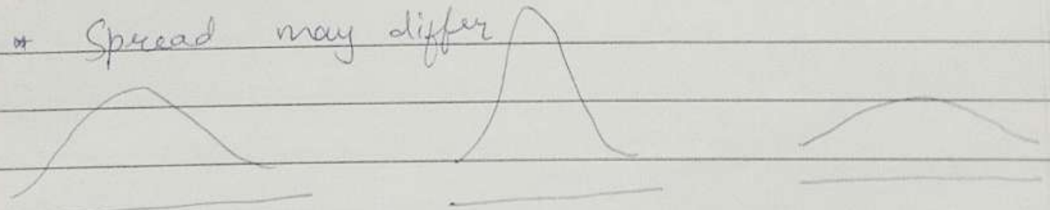


Kernel density estimator

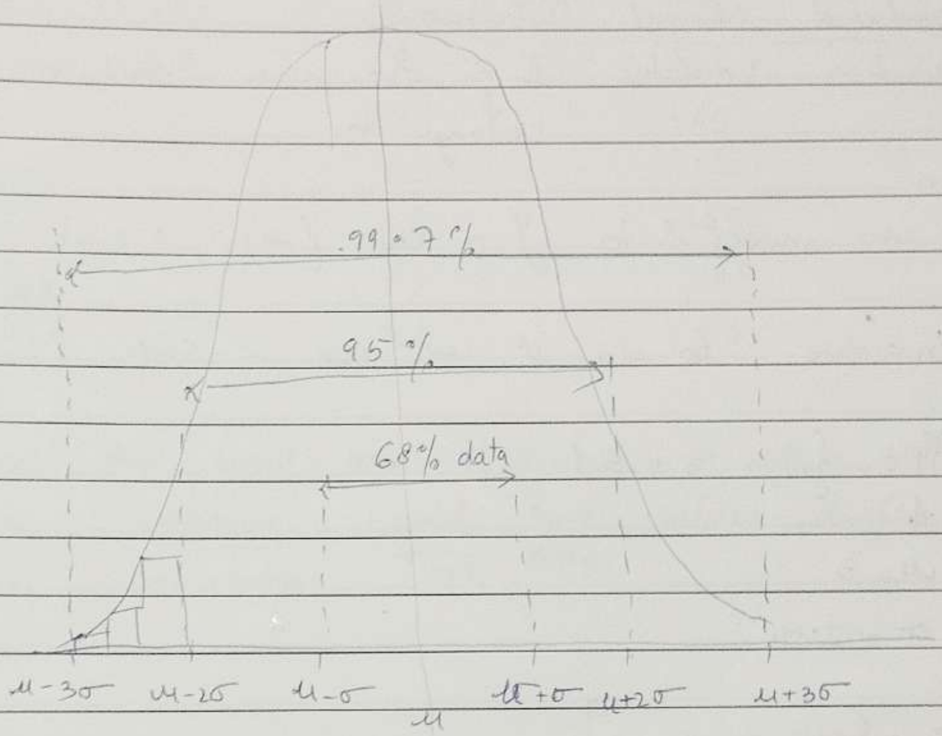
Age, weight, height all follow Gaussian/ Normal distribution - {Domain Expertise say this}

IRIS DATASET → Also follow Gaussian/Normal distribution

Note * Spread may differ



* [Emperical Rule of Normal Distribution]



* Using Normal/Standard Deviation we can make assumptions o
the data.

1) In a distribution {dataset} if it follows Gaussian distribution:
i) 68% will be embedded within the first standard deviation to th
left and first standard deviation to the right.
ii) 95% will be within the second standard deviation to the rig
& left.
iii) 99.7% of data will be within the third standar deviation to
left & right.

→ This is Emperical formula

$\boxed{68\% - 95\% - 99.7\%}$ Rule

RANKA
DATE / /
PAGE

(Q-Q) plot we can check if distribution is Gaussian or not.

* Standard Normal Distribution
Assume Random Variable $X \approx$ Gaussian Distribution with mean (μ, σ)
belongs to

then,
we can convert X to $y \approx$ SND $(\mu = 0, \sigma = 1)$

→ Inorder to do it we use Z-Score.

= $X = \{1, 2, 3, 4, 5\}$
↓
$\mu = 3$
$\sigma = 1.41$

$$Z\text{-Score} = \frac{x_i - \mu}{\boxed{\frac{\sigma}{\sqrt{n}}}}$$ ⇒ Standard Error ⇒ Useful in Inferential stats

sample data 4
$\boxed{n = 1}$ ⇒ because we apply it to every variable

if $n = 1$ → Sample size

$$\boxed{Z\text{-score} = \frac{x_i - \mu}{\sigma}}$$

$X = \{1, 2, 3, 4, 5\}$
$\mu = 3$ $\sigma = 1.414$

$z = \frac{1-3}{1.414} = -1.414$

$= \frac{2-3}{1.414}$

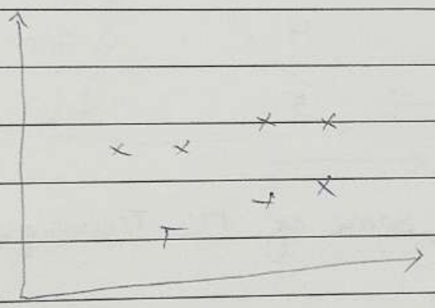
→ $y = \{-1.414, -0.707, 0, 0.707, 1.414\}$
↑
$\mu = 0, \sigma = 1$

Note :- Assumptions of Gaussian Distribution will be also applied here.

Q. Why we are converting to Standard Normal distribution.

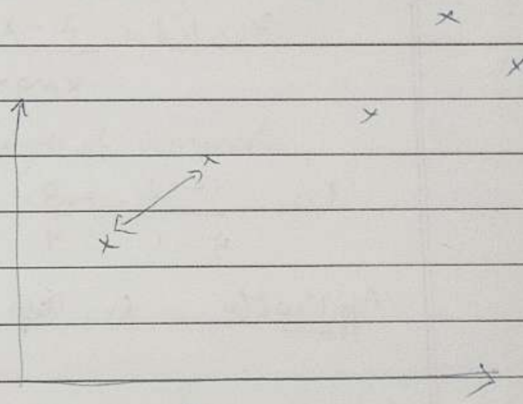
(Years) Age	(kg) Weight	(cm) Height
24	72	150
26	78	160
32	84	165
33	92	170
34	87	150
28	83	180
29	80	175

* Here all the units are different, so when we make a Machine learning model these values differ by a lot, therefore calculation Time ↑↑↑.



I

SND



II

→ In case of the I graph the calculation time would be less in case of Euclidean distance.

→ With standard scaling we will bring it to same scale

→ This entire process is called standardization ⇒ $\boxed{\mu = 0 \ \& \ \sigma = 1}$

NORMALIZATION $\Rightarrow$ Assumptions of Gaussian distribution not applied.
In case of normalization we convert values into specific set of range.

[0,-1] [-1 - 1]

[0-5]

[0- 4]

In normalization ~ [lower Scale $\leftrightarrow$ Higher Scale]

(1) In Min Max Scaler [0 - 1]

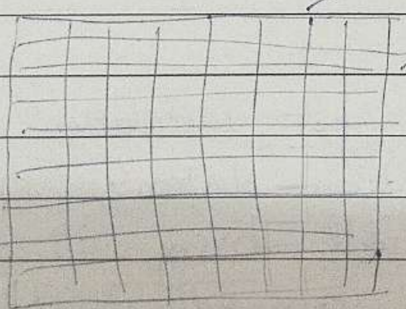
$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

$$x_1 = \frac{1-1}{5-1} = \frac{0}{4}$$

x	$\Rightarrow y$
1	0
2	0.25
3	0.5
4	0.75
5	1

Application - In Deep learning, some of ML Techniques.

→ Images → Pixel



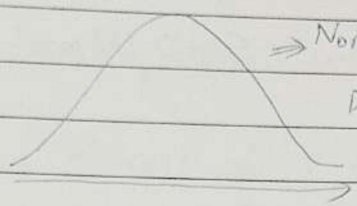
→ [0 - 255] pixels $\leftrightarrow$ [0-1]

Notes :- All of these techniques are feature scaling

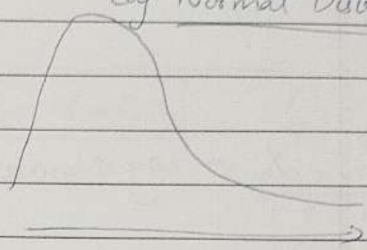
* We convert to SND → Bring the features in the same scale.

LOG NORMAL DISTRIBUTION

Log Normal Distribution

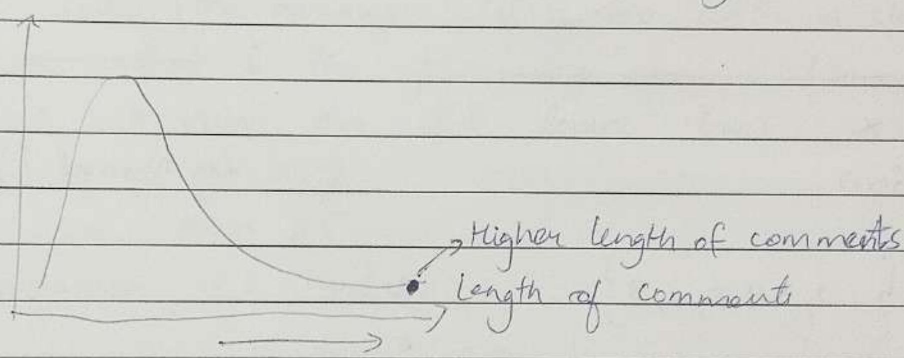


⇒ Normal/Gaussian Distribution

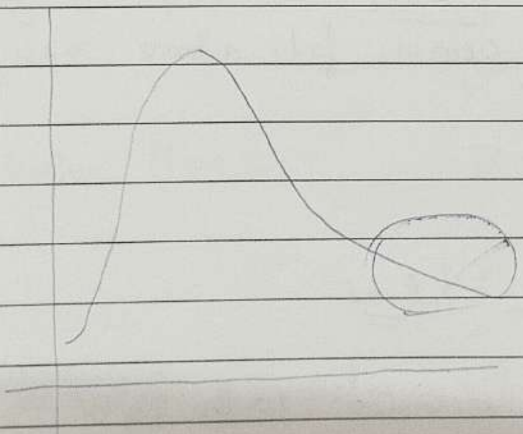


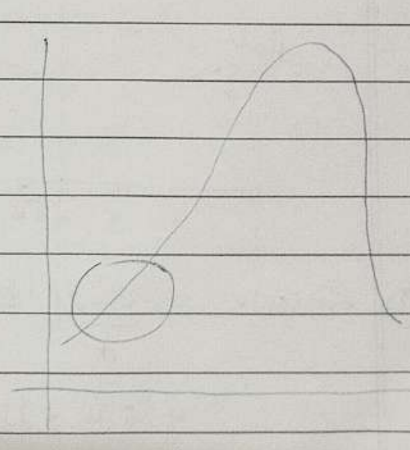
This is right skewed

Eg:- Wealth Distribution, Youtube comments length.



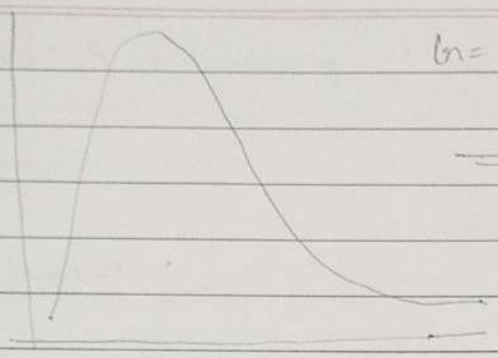
Q.





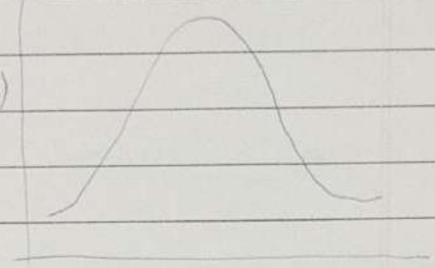
From Ascending order give the relationship of mean, median & mode → $\boxed{\text{Clue → look at the outliers}}$

*



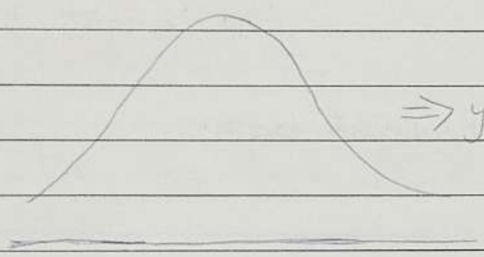
ln = $\log_e$ = natural log

$\Longrightarrow Y = \ln(X)$

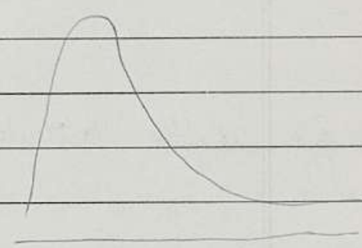


SND (Normal Distribution)

X > log Normal Distribution



$\Rightarrow y = \exp(x) \Rightarrow$



X

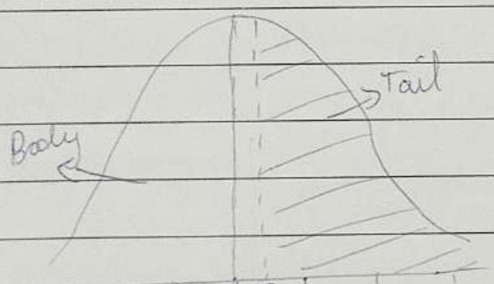
SND

Log Normal Distribution

* $X = \{1, 2, 3, 4, 5, 6, 7\}$

$\mu = 4$

$\sigma = 1$



Question: What percentage of scores fall above 4.25%

① Z-score $= \dfrac{x - \mu}{\sigma} = \dfrac{4.25 - 4}{1} = 0.25$

4.25 is 0.25 standard deviation to the right

* Z-table (area under the curve)

Z-score for 0.25 = 0.5987

∴ Area of Tail = 1 − 0.5987 = 40.13%

RANKA
DATE / /
PAGE

ii) Percentage of score that falls below 3.75.

Ans = $Z\,score = \frac{3.75 - 4}{1} = -0.25 \rightarrow -0.25$ standard deviation to the left

Area is 40.13

iii) Area of the curve between $\boxed{4.75 \;\&\; 5.75}$

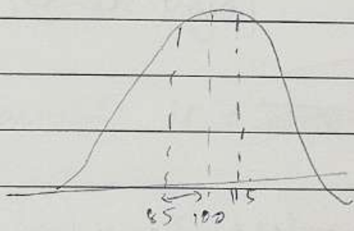
$$\begin{array}{r} 95.99 \\ -\;77.34 \\ \hline 18.65 \end{array}$$

Z score of 0.75 & 1.75

$77.34 - 95.99 = 18.65\%$

Q. In India the average IQ is 100 with a standard deviation of 15. What is the percentage of population would you expect to have an IQ lower than
i) lower than 85%
ii) higher than 85
iii) between 85 and 100

Ans. i) $\mu = 100$, $\sigma = 15$



$Z\,score = \frac{x - \mu}{\sigma} = \frac{-15}{15} = -1$

$= .1587$

ii) higher than 85, Z-score $= -1$

$1 - .1587 = 84.13\%$

iii) between 85 and 100 it would be $.3413$ or 34.13%